LAPORAN PENELITIAN

ITEM ANALISIS, STATISTIKA DAN PENURUNAN RUMUS ITEM ANALISIS

Oleh :

Dra. Nani Dianiyati
Drs. Herman . MA

Universitas Terbuka
Lembaga Penelitian
Pusat Studi Indonesia
1997

Lembar Pengesahan
Laporan Penelitian PSI-UT

1. a. Judul Penelitian : Item Analisis, Statistika dan Penurunan Rumus Item Analisis
   b. Bidang Penelitian : Statistika

2. Ketua Peneliti
   a. Nama lengkap dan gelar : Dra. Nani Dianiyati
   b. NIP : 131-627-868
   c. Golongan kepangkatan : III/c
   d. Jabatan Fungsional : Lektor Muda
   e. Fakultas/Unit Kerja : FMIPA/Pusat Komputer

3. Anggota tim peneliti
   a. Jumlah anggota : 1 orang
   b. Nama anggota/NIP/Gol. Kepangkatan :
      1. Herman/131-628-379/III/c

4. Lama Penelitian : 6 bulan

5. Biaya Penelitian : Rp.3.516.000,-
   (Tiga juta lima ratus enam belas ribu rupiah).

Pondok Cabe, 4 Agustus 1997

Mengetahui,
Dekan FMIPA

Dr. Djati Kerami
NIP 130 422 587

Ketua Peneliti

Dra. Nani Dianiyati
NIP 131 627 868

Menyetujui,
Kepala PSI-UT

Dr. Tian Belawati
NIP 131 569 974

Menyetujui,
Ketua Lembaga Penelitian UT

WBP Simanjuntak, MEd, PhD
NIP 130 212 017

Abstrak

Pada penelitian ini dijelaskan konsep konsep yang digunakan untuk item analisis yang biasa dikenakan pada soal-soal pilihan ganda. Di item analisis konsep-konsep yang yang dipakai adalah konsep matematik dan statistik serta konsep dari teori psikologi yang diterapkan di pendidikan. Karena itu penelitian ini mengaitkan hal hal tersebut di atas.

Konsep-konsep psikologi yang dipakai pada pengukuran kemampuan (achievement) pelajar/mahasiswa seperti teori tentang *true score* diuraikan disini. Selain itu, konsep konsep statistik seperti mean, variansi, kovariansi dan korelasi yang digunakan untuk mengukur reliabiliti (keandalan) suatu set soal ataupun untuk menghitung besaran-besaran item analisis lainnya juga diuraikan pada penelitian ini.

Daftar Isi

"ITEM ANALISIS, STATISTIKA DAN PENURUNAN RUMUS ITEM ANALISIS".

I. Pendahuluan

I.1 Latar Belakang

Untuk menguji atau mengetahui tingkat pemahaman siswa dan mahasiswa terhadap materi yang diberikan oleh guru / dosen, salah satu caranya adalah dengan memberikan soal-soal ujian. Soal-soal ujian tersebut haruslah soal-soal yang "baik". Artinya soal-soal tersebut harus dapat mengukur penguasaan materi (yang sesuai dengan kurikulum) yang harus dimiliki oleh siswa / mahasiswa. Agar soal yang akan diberikan pada ujian mempunyai kualitas "baik", soal-soal tersebut haruslah dibuat oleh pakar pada mata pelajaran / mata kuliah yang akan diujikan. Karena itu, kumpulan soal-soal (set soal) yang akan diujikan diharapkan benar-benar dapat mencerminkan materi yang harus dikuasai oleh siswa / mahasiswa.

Kata "baik" di sini dipandang dari sudut si pembuat soal sebelum soal tersebut diujikan. Persoalannya sekarang, "apakah baik dari sisi si pembuat soal juga akan baik bila dilihat dari sisi si siswa yang diuji?" Untuk menjawab pertanyaan ini, soal soal tersebut haruslah diuji secara statistik dengan alat yang disebut item analisis.

Item analisis adalah suatu alat yang sering dipakai di dunia pendidikan. Guna alat ini adalah untuk melihat seberapa "baik" soal-soal yang diberikan kepada siswa / mahasiswa. Dalam menguji mutu soal-soal tersebut ada beberapa besaran yang diukur. Besaran-besaran inilah yang nantinya akan mencerminkan "mutu" dari soal-soal yang diujikan.

Besaran-besaran yang ada pada item analisis pada umumnya adalah tingkat kesulitan, nilai rata-rata siswa untuk satu set soal, nilai rata-rata siswa per nomor soal, KR-20, K-bis, standard deviasi nilai peserta ujian, dan beberapa besaran lainnya. Penelitian ini mencoba menjelaskan bagaimana besaran-besaran tersebut dapat dikaitkan dengan "kemampuan soal" yang mengukur kemampuan mahasiswa dalam menguasai materi pelajaran. Penurunan rumus-rumus yang digunakan untuk menghitung besaran-besaran tersebut diturunkan secara matematik.

I.2 Anggapan

Agar set soal yang diberikan kepada siswa / mahasiswa benar-benar mencerminkan isi kurikulum, ada beberapa hal yang harus dipenuhi. Yang pertama adalah soal-soal harus dibuat oleh pakar mata pelajaran (mata kuliah) yang sesuai. Karena sebagai seorang pakar, ia dianggap tahu jenis soal yang harus diberikan untuk menguji kemampuan siswa dalam penguasaan materi pelajaran.

Untuk membantu pakar-pakar tersebut dalam membuat soal, ada suatu alat yang dapat dipakai. Alat tersebut adalah kisi-kisi soal. Kisi-kisi soal ini dibuat oleh pakar-pakar, sesuai bidang ilmu masing-masing. Alat ini dianggap dapat mencerminkan isi kurikulum. Karena itu, hal kedua yang harus dipenuhi adalah "pembuatan set soal haruslah sesuai dengan kisi-kisi soal".

I.3 Perumusan Masalah

Pada pemakaian item analisis terdapat beberapa formula statistik yang dipakai. Formula ini terdapat di buku-buku teks yang ada. Akan tetapi pada literatur hampir tidak ada penjelasan asal muasal formula yang dipakai secara matematika. Bagaimana sebetulnya asal mula terjadinya formula tersebut ? Disamping itu latar belakang teori teori psikologi yang dipakai pada item analisis tidak banyak dijelaskan. Padahal teori ini penting untuk lebih memperjelas "cara bekerja"nya item analisis.

I.4 Tujuan Penelitian

Penelitian ini bertujuan untuk menjabarkan formula-formula Statistik yang dipakai pada item analisis. Dalam penjabaran tersebut dikupas juga teori teori psikologi yang dipakai di dunia pendidikan (khususnya pada item analisis)

I.5 Manfaat Penelitian

Penelitian ini mempunyai beberapa manfaat. Pertama, hasil penelitian dapat digunakan untuk menilai seberapa "baik" mutu soal-soal dari tes objektif yang ditawarkan kepada siswa/mahasiswa di Indonesia ini.

Manfaat yang kedua adalah menambah pengalaman penelitian untuk si peneliti. Dengan demikian, untuk penelitian selanjutnya diharapkan kualitas penelitian akan semakin baik. Pengetahuan dan wawasan si penelitipun diharapkan akan bertambah banyak dan luas.

Pengembangan ilmu statistika dan matematika di kependidikan adalah manfaat yang ke tiga. Diharapkan untuk jangka panjang penelitian yang mengaitkan ilmu statistika, matematika dan kependidikan dapat terus dilakukan.

II. Metode Penelitian

1. Studi literatur tentang teori psikologi yang digunakan untuk pendidikan.
2. Studi literatur tentang teori-teori statistik yang biasa dipakai didalam behavioral science, termasuk kependidikan.
3. Menghubungkan teori psikologi di kependidikan dengan teori statistik.
4. Penurunan formula statistik yang digunakan untuk item analisis.
5. Pembuatan laporan.

III. Tinjauan Pustaka

Menurut Crocker dan Algina (1986), item analisis tidak dapat dilepaskan dari teori tes. Pada teori tes ada dua metode yang dipelajari, yaitu 1) menaksir sampai dimanakah problem-problem yang ada akan mempengaruhi pengukuran pada situasi tertentu, 2) membuat metode metode untuk mengatasi atau meminimalkan problem-problem yang muncul. Sedangkan teori tes di dalam pemakaiannya tidak dapat dipisahkan dengan logika dan model matematika. Untuk mengerti tentang teori tes ada beberapa konsep dasar yang harus diketahui. Konsep konsep tersebut adalah tentang pengukuran (measurement), konstruk (construct), dan tes psikologi (psychological test).

Weitzenhoffer (1951) mendifinisikan measurement sebagai suatu operasi yang dilakukan oleh peneliti di dunia fisik. Sedangkan Stevens (1946) mendifinisikan measurement sebagai angka-angka yang diberikan ke pada objek-objek atau kejadian-kejadian berdasarkan kaidah-kaidah tertentu. Lord dan Novick (1968), juga Torgerson (1958) melengkapi difinisi Stevens tentang measurement dengan menambahkan bahwa pengukuran tersebut lebih digunakan pada sifat-sifat objek daripada terhadap objek itu sendiri.

Bila seorang ahli kimia mengukur berat molekul suatu atau seorang ahli biologi menghitung jumlah bakteri di suatu tempat maka mereka melakukan pengukuran terhadap atribut tertentu dari objek yang mereka teliti. Analogi dengan contoh di atas, seorang ahli jiwa anak-anak tidaklah mengukur anak-anak tersebut, tetapi mereka mengukur atribut yang dimiliki oleh anak-anak itu. Atribut-atribut ini dapat berupa atribut fisik seperti tinggidan berat badan ataupun atribut psikologi seperti perkembangan kosa-kata, kemampuan sosialisasi, dan pengetahuan-pengetahuan tertentu.

Secara traditional psikologi didifinisikan sebagai studi tentang behavior (kelakuan) dan atribut yang mencirikan kelakukan seseorang di rumah di kantor ataupun di sekolah dan pada interaksi sosial. Atribut atribut ini dikenal sebagai *psychological attributes* (beberapa penulis mengistilahkan dengan *psychological traits*).

Tidak seperti atribut fisik, atribut psikologi tidak langsung dapat diukur seperti mengukur tinggi dan berat badan. Atribut psikologi adalah *construct*. Construct adalah konsep hipotetis yang dihasilkan oleh ahli-ahli ilmu sosial dari khayalan scientifik yang bertujuan untuk mengembangkan teori yang dapat menjelaskan human behavior. Menurut Crocker dan Algina (1986) keberadaan construct tidak dapat dipastikan secara mutlak. Karena itu derajat *psychological construct* yang mencirikan behavior seseorang hanya dapat diduga dari observasi terhadap kelakuan orang tersebut. Ada baiknya ilustrasi berikut ini diperhatikan untuk memperjelas bagaimana pengukuran suatu *construct*.

Seorang ahli jiwa memperhatikan sekelompok anak-anak yang sedang bermain. Ia melihat bahwa beberapa anak-anak sering sekali mengarahkan teman-temannya untuk berbuat sesuatu. Karena kejadian ini ia lihat berulang kali maka ia mengistilah kelakuan beberapa orang anak-anak tadi sebagai pendominasian sosial (*socially dominating*). Disini ahli jiwa menemukan apa yang disebut dengan *construct* (teoritis). Tetapi harus diingat bahwa menemukan *construct* tidaklah sama dengan mengukur *construct*.

Sebelum pengukuran terhadap *construct* (praktis) dapat dilakukan maka ada beberapa aturan- yang berkaitan dengan *construct* teoritis dan juga berkaitan dengan kelakuan yang dapat diobservasi- yang harus dibuat. Dengan kata lain harus dibuat difinisi construct yang lebih operasional sehingga *construct* tadi dapat diukur. Setelah itu alat untuk mengukur barulah dapat dibuat. Alat inilah yang dikenal dengan istilah ***tes***.

Tes dapat didifinisikan sebagai "prosedur standar untuk membuat suatu sampel dari behavior dari domain tertentu". Karena itu, tes berkaitan dengan prosedur untuk membuat sampel dari penampilan *optimal* individu individu (contoh pada ujian individu ini haruslah menjawab sebaik mungkin) dan sampel dari penampilan *tipikal* (contoh sewaktu mengisi kuesioner individu ini harus menjawab sejujur mungkin).

Untuk item analisis, difinisi operasional dari *construct* yang akan diukur antara lain adalah reliabiliti soal, derajat kesulitan soal, daya beda soal, nilai rata-rata peserta tes dan variansi nilai peserta tes. Nilai disini adalah *raw score* (nilai mentah). Besaran besaran ini berkaitan dengan konsep konsep matematika.

Reliabiliti gunanya untuk melihat seberapa konsisten soal-soal itu dijawab oleh siswa. Kalau saja jawaban siswa tersebar dalam banyak variasi maka soal tersebut tidaklah reliabel. Misalnya ada empat stem jawaban yaitu a, b, c dan d. Kalau untuk soal no. 1 sebagian besar siswa menjawab ke salah satu jawaban, sedangkan pada soal no. 2 jawaban siswa tersebar merata pada ke empat stem jawaban maka soal no.1 lebih reliable dari pada soal no. 2. Koefisien reliabiliti bergerak dari nilai 0 ke 1.

Derajat kesulitan suatu soal berguna untuk melihat berapa banyak siswa yang menjawab benar soal tersebut. Semakin banyak yang bisa menjawab benar maka semakin mudah soal itu. Demikian juga sebaliknya, semakin sedikit yang bisa menjawab benar maka semakin sulit soal itu. Nilai derajat kesulitan bergerak dari 0 ke 1. Semakin besar nilai derajat kesulitan maka semakin mudahlah soal itu.

Guna daya beda soal adalah untuk melihat seberapa mampukah soal ini membedakan siswa yang pandai dengan siswa yang tidak pandai. Kalau daya beda soal ini "bagus", maka soal ini akan mampu membedakan mana siswa yang pandai dan mana siswa yang tidak pandai.

Nilai rata-rata berguna untuk melihat pada posisi manakah rata-rata nilai siswa berada diantara kontinum nilai minimum dan maksimum. Sedangkan nilai variansi gunanya untuk melihat seberapa besar sebaran data nilai-nilai siswa tersebut. Semakin besar nilai variansi semakin menyebar pula nilai siswa peserta ujian. Semakin kecil nilai variansi maka nilai siswa akan semakin menumpuk disuatu daerah nilai.

IV. Pembahasan

IV.1 Konsep Statistika pada Item Analisis

Menurut Crocker & Algina (1986) ada beberapa konsep statistik yang banyak digunakan pada teori tes. Konsep konsep tersebut adalah pengertian variabel (diskrit dan kontinu), ukuran keterpusatan (measures of central tendency), ukuran keberagaman (measures of variability), distribusi normal, distribusi binomial, serta korelasi dan regresi.

Dalam meneliti suatu populasi, karakteristik diskriptif dari individual dapat *konstan* (tetap) dan dapat juga *berubah* (variabel). Konstan adalah suatu *karakteristik yang sama* yang dimiliki oleh setiap anggota populasi. Contoh nilai konstan adalah suatu penelitian hanya melibatkan siswa SMU. Sehingga anggota populasi tersebut adalah siswa SMU saja.

Variabel adalah suatu *karateristik* yang dimiliki oleh setiap anggota populasi *yang nilainya dapat berbeda* satu dengan yang lainnya. Contoh variabel adalah berat. Setiap anggota populasi akan mempunyai berat. Tetapi nilai berat masing-masing individu belum tentu sama. Di dalam penelitian bilamana populasinya berhingga dan nilai-nilai variabelnya berbentuk bilangan bulat maka ia biasanya disebut dengan *variabel diskrit*. Sebaliknya bilamana populasinya tak berhingga (unlimited) dan nilai variabelnya tidak hanya berupa bilangan bulat, tetapi ada bilangan pecahan diantara 2 bilangan bulat, maka variabel tersebut biasanya dinamakan *variabel kontinu*.

Ukuran keterpusatan dapat menunjukkan posisi skor individu di dalam grup. Untuk ukuran keterpusatan ini ada tiga jenis yang dipakai, yaitu : mean (nilai rata-rata), median (nilai tengah) dan mode. Mean dari suatu obsevasi adalah nilai rata-rata dari semua nilai observasi tersebut. Formula yang dipakai adalah :

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}, \quad i = 1, 2, 3, \ldots\ldots, N$$

Median adalah titik yang berada di tengah-tengah dari distribusi nilai pengamatan. Pada pengamatan, nilai-nilai pengamatannya diurutkan terlebih dahulu. Kalau banyaknya observasi adalah ganjil, maka observasi yang ke (n+1)/2 adalah nilai tengahnya. Seandainya banyaknya observasi tersebut genap maka nilai tengahnya adalah setengah dari jumlah nilai observasi yang ke n/2 dan nilai dari observasi yang ke (n/2) + 1. Akan tetapi pada prinsipnya titik tengah menunjukkan 50% skor akan berada dibawahnya dan 50% lagi akan berada diatasnya.

Mode adalah nilai pengamatan yang paling banyak muncul. Istilah lainnya mode sama dengan nilai frekuensi dari pengamatan yang tertinggi. Kalau suatu pengamatan mempunyai beberapa nilai yang frekuensinya sama maka modenya akan lebih dari satu.

Ukuran keterpusatan masih belum cukup untuk menjelaskan keadaan dari grup yang diteliti. Untuk itu dibuatlah ukuran variabiliti observasi. Ada tiga macam ukuran variabiliti yang dikenal, yaitu *range*, *variansi* dan *standar deviasi*.

Range biasanya akan memperlihatkan nilai minimum dan maksimum yang ada pada observasi. Selisih dari nilai maksimum dan minimum itulah yang dikenal dengan *range*. Sedangkan variansi dan standar deviasi dibuat berdasarkan selisih nilai (dari tiap observasi terhadap nilai mean). Formula untuk variansi akan dibedakan antara populasi (σ^2)dan sampel(S^2). Pada sampel nilai penyebutnya adalah (N - 1). Alasannya karena bentuk tersebut adalah bentuk penaksir unbiased (Hoog & Craig, 1978).

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}, \qquad i = 1, 2, 3, \ldots, N$$

$$S^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}, \qquad i = 1, 2, 3, \ldots, N$$

Nilai standar deviasi adalah adalah akar dari nilai variansi. Dari nilai variansi ataupun nilai standar deviasi dapat dilihat seberapa besar variasi data observasi yang ada. Semakin kecil nilainya semakin kurang bervariasi pula datanya. Sebagai ilustrasi misalnya nilai rata-rata suatu mata kuliah pada suatu observasi adalah 60 (nilai minimum 0 dan nilai maksimum 100). Variansinya adalah 10. Ini artinya data observasi yang berada antara 50 dan 70 besarnya 65% dari keseluruhan data. Tetapi kalau variansinya adalah 30, maka 65% data akan berada pada interval 30 dan 90. Sehingga data dengan variansi 30 akan lebih bervariasi dibandingkan data dengan variansi 10.

Kumpulan data pada observasi akan membentuk suatu distribusi. Pada penelitian ini ada dua distribusi yang akan dibahas. Distribusi tersebut adalah distribusi binomial dan distribusi normal. Distribusi binomial dipakai pada suatu eksperimen dimana hasil percobaan tersebut dapat digolongkan dalam dua kejadian yang saling bebas (mutually exclusive and exhaustive ways).

Bentuk persamaan distribusi tersebut adalah :

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & , x = 1, 2, \ldots\ldots, n \\ 0 & , x \text{ lainnya} \end{cases}$$

dimana $0 < p < 1$.

Distribusi normal adalah suatu distribusi yang simetri terhadap sumbu tegak. Keistimewaannya adalah nilai mean, median dan modus berada pada satu titik yang sama. Distribusi ini mempunyai bentuk :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{x-\mu}{2\sigma^2}\right] \quad , \; -\sim < x < \sim$$

Distribusi normal ini sering dipakai di dalam penelitian. Apalagi ada teorema limit sentral yang menyatakan distribusi dari nilai rata-rata akan berdistribusi normal.

Konsep lain yang digunakan adalah korelasi. Korelasi adalah hubungan linear antara 2 variabel. Besar kecilnya hubungan itu dapat dilihat dari koefisien korelasi yang nilainya berada antara -1 dan 1. Tanda minus menunjukkan bahwa jika suatu variabel nilainya naik/turun, maka nilai variabel lainnya akan turun/naik. Sedangkan tanda plus menyatakan jika suatu variabel nilainya naik/turun maka nilai variabel lainnya akan naik/turun juga. Korelasi antara variabel X dan Y didifinisikan sebagai :

$$\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

dimana $Cov(X,Y) = \dfrac{\sum_{i=1}^{N} (X_i - \mu_x)(Y_i - \mu_y)}{N}$

Selain konsep-konsep di atas ada beberapa konsep lain yang juga dipakai. Untuk itu, ilustrasi berikut ini perlu diperhatikan. Sebuah uang logam dilempar (toss), dan misalkan hasilnya adalah muka (H) atau belakang (T). Seandainya toss uang logam tersebut dapat diulangi dengan kondisi yang sama, maka toss dari mata uang ini adalah satu contoh dari **eksperimen random** dimana hasil keluarannya adalah H atau T. Pada percobaan ini H dan T disebut sebagai **ruang sampel** dari eksperimen.

Misalkan *C* menunjukkan ruang sampel dan misalkan C menunjukkan bagian dari *C*. Bila pada percobaan hasilnya berada di dalam C, maka *event* C terjadi. Sekarang misalkan ada N percobaan random. Frekuensi terjadinya event C katakanlah sama dengan *f* kali. Rasio *f/N* disebut sebagai *frekensi relatif* event

C dalam N eksperimen. Frekuensi relatif untuk N yang "kecil" tidaklah terlampau menarik perhatian. Tetapi untuk N yang semakin besar, tampak bahwa frekuensi akan semakin stabil. Biasanya frekuensi ini akan disimbolkan sebagai $p = f/N$. Walaupun hasil keluaran suatu eksperimen tidak dapat diramalkan, tetapi untuk N yang besar frekuensi relatif dapat diramalkan. Bilangan ***p*** yang dikaitkan dengan event C dikenal dengan istilah ***peluang***.

Himpunan-himpunan aljabar adalah konsep lain yang juga perlu diketahui. Biasanya konsep himpunan atau koleksi dari objek-objek dibiarkan saja tanpa penjelasan. Akan tetapi *suatu himpunan tertentu* dapat menjelaskan arti dari himpunan tersebut tanpa adanya salah pengertian tentang koleksi objek-objek yang sedang dibahas. Misalnya himpunan 10 bilangan bulat positif yang lebih kecil dari 11. Dari sini tampak jelas bahwa angka 11 dan ½ tidaklah termasuk unsur himpunan itu. Khusus untuk penelitian ini himpunan yang dipakai adalah himpunan bilangan real.

IV.2 Penskalaan

Dari difinisi pengukuran, penomoran pada sifat-sifat objek harus dibuat berdasarkan aturan-aturan tertentu. Pengembangan aturan-aturan itu secara sistimatis serta pendifinisian unit ukuran untuk mendifinisikan simbol simbol angka pada objek yang diobservasi disebut dengan *penskalaan* (scalling). Aturan pada penskalaan akan membentuk kaitan dari suatu sistim data dengan elemen elemen dalam sistim bilangan real.

Stevens (1946) mendifinisikan empat macam penskalaan yang berbeda. Ke empat jenis penskalaan itu adalah **nominal**, **ordinal**, **interval** dan **ratio**. Skala nominal dipakai untuk memberikan nomor kepada objek-objek. Nomor-nomor ini tidak mempunyai arti *selain dari* identitas suatu objek, jadi tidak ada arti urutan nomor ataupun tidak ada arti jarak antara nomor. Misalnya nomor punggung pada pemain bola. Disini pemberian nomor tidak ada kaitannya dengan kemampuan. Artinya kemampuan bermain bola pemain nomor 1 belum tentu lebih buruk/baik dari kemampuan pemain yang mempunyai nomor punggung 10. Pada penelitian sosial skala ini biasanya dipakai untuk memberi identitas pada jenis kelamin, jenis pekerjaan, agama, ataupun untuk menunjukkan daerah tempat. Kode UPBJJ yang dimiliki oleh Universitas Terbuka adalah salah satu contoh skala nominal.

Bilamana elemen nomor yang diberikan pada data dapat diurutkan berdasarkan sifat-sifat yang diukur maka penskalaan yang digunakan adalah skala ordinal. Tidak seperti pada skala nominal, pada skala ordinal *urutan* mempunyai arti. Salah satu skala ordinal yang terkenal adalah yang dibuat oleh Likert (1932). Ia membagi satu garis kontinum menjadi lima bagian yaitu strongly disagree (1), disagree (2), neutral (3), agree (4), strongly agree (5). Tetapi pada skala ini jarak antara satu titik dengan titik yang lainnya belum tentu sama. Jadi jarak antara strongly disagree dengan disagree belum tentu sama dengan jarak dari disagree dengan neutral misalnya.

Apabila jarak antara satu titik dengan titik yang lainnya dapat dibandingkan dan juga urutan juga menjadi perhatian maka skala yang dipakai adalah skala interval. Disini urutan data ada dan apabila beberapa titik mempunyai selisih jarak yang sama, maka selisih jarak mereka juga sama terhadap sifat-sifat yang diukur. Contoh yang digunakan oleh Crocker & Algina (1986) untuk menjelaskan skala interval adalah pengukuran panjang benda dengan menggunakan kartu bridge. Benda yang satu berukuran 10 kartu bridge sedangkan benda-benda yang lain berukuran 11, 23 dan 24 kartu bridge. Disini selisih panjang benda yang mempunyai ukuran 10 dan 11 kartu bridge akan sama dengan selisih panjang benda yang berukuran 23 dan 24 kartu bridge.

Penskalaan yang ke empat adalah skala ratio. Skala ratio disamping memiliki sifat dari ke tiga skala yang lainnya, ia juga mempunyai nilai 0 absolut. Temperatur 0 bukanlah nilai 0 absolut karena 0 pada Celcius dan Fahrenhit tidaklah sama. Nol pada temperatur menunjukkan nilai dari tempratur tersebut. Tetapi 0 absolut lebih berarti sebagai titik asal (fixed origin). Skala ratio banyak digunakan pada fisika, misalnya panjang dalam cm, berat dalam gram. Sekali lokasi nilai 0 absolut diketahui, maka ukuran bukan 0 pada skala ini dapat ditentukan sebagai ratio dari dari satu ukuran dengan yang lainnya.

IV.3 Pendekatan Penskalaan untuk Pengembangan Tes

Pada pengembangan instrumen untuk pendidikan ataupun psikologi, pengembang tes biasanya akan terlibat dalam serangkaian pengujian hipotesa tentang kemampuan skala (scalability) dari data pengukuran untuk suatu construct. Yang pertama adalah membuat hipotesa dari data yang berasal dari construct yang sudah diukur. Ukuran (operasional) ini ditempatkan pada suatu continuum (biasa dikenal dengan psychological continuum). Kedua adalah mencari jawab pertanyaan tentang sifat-sifat bilangan real yang manakah yang dimiliki oleh skala yang ada pada continuum. Torgerson (1958) memperkenalkan 3 macam pendekatan untuk meneliti ke dua hal diatas, yaitu *subject-centered methods, stimulus-centered methods,* dan *response-centered methods*.

Pendekatan dengan metode subject-centered digunakan untuk melihat posisi beradanya individu-individu pada titik-titik di continuum. Pengukuran sikap (aptitude) dan prestasi belajar (achievement) biasanya menggunakan pendekatan ini. Skala ordinalpun akan dapat dipakai disini karena skala ini dapat menempatkan individu-individu pada titik-titik di continuum. Tetapi skala nominal tidak dapat digunakan disini.

Stimulus-centered methods lebih menekankan pada penempatan stimuli (item) itu sendiri di dalam continuum. Jadi tempat dimana posisi stimuli (item) tersebut berada pada continuum, adalah tujuan pemakaian metoda ini. Misalkan satu soal pada ujian dianggap sebagai "soal yang mudah" oleh si pembuat soal, tetapi setelah diujikan kepada banyak orang ternyata soal tersebut berada pada posisi "soal dengan derajat kesulitan sedang". Jadi respons terhadap suatu

stimuli akan menghasilkan penempatan posisi stimuli tersebut pada suatu continuum.

Responce-centered approach adalah pendekatan yang paling rumit diantara ke tiga pendekatan di atas. Respon terhadap data digunakan untuk memberi skala pada subjek-subjek dan meletakkannya pada continuum berdasarkan jawaban yang benar. Bersamaan dengan itu item-itemnya pun diberi skala dan diletakkan pada continuum.

IV.4 Statistik Deskriptif untuk Variabel Nondikotomis / dikotomis

Statistika, seperti yang dapat dibaca pada buku-buku teks dapat digunakan untuk dua hal yaitu: 1) menjelaskan distribusi nilai-nilai dari data observasi (statistika deskriptif) dan 2) menaksir "keadaan" suatu populasi -dengan uji hipotesis- dari suatu sampel (statistika inferensi).

Misalkan ada 10 siswa yang mengikuti ujian. Ujian tersebut terdiri dari 5 soal, dimana masing-masing soal mempunyai nilai minimum 0 dan nilai maksimum 5. Dari informasi ini tempat siswa didalam continuum akan berada diantara nilai 0 dan 25. Siswa-siswa yang menjawab benar semua akan berada pada titik 25, dan siswa-siswa yang tidak mempunyai jawaban benar satu soalpun akan berada pada titik 0. Sementara itu siswa-siswa yang lainnya akan berada diantara titik 0 dan 25. Selain itu posisi item (stimuli) pun dapat ditentukan. Soal nomor 1 sampai dengan nomor 5 akan berada pada continuum 0 dan 50. Item yang dapat dijawab benar oleh setiap peserta tes akan berada pada titik 50. Sedangkan item yang tidak dapat dijawab oleh setiap peserta tes akan berada pada titik 0. Sementara itu item-item lainnya akan berada di antara titik 0 dan 50.

Item (stimuli) inipun dapat dibuat korelasinya. Tiap item, mulai dari item 1 sampai dengan item 25 akan memiliki jawaban dari masing-masing peserta. Jawaban jawaban ini lalu dibandingkan untuk masing-masing item. Hasil dari perbandingan ini disebut korelasi. Jadi kalau item 1 dan item 2 jawabannya persis sama, maka korelasi ke dua item tersebut akan sama dengan 1.

Hasil perhitungan, posisi peserta tes/item yang ke akan berada pada :

$$\text{posisi peserta ke-j} = \sum_{i=1}^{5} x_{ij}, \quad x_{ij} \text{ mempunyai nilai dari } 0 \text{ sampai } 5$$

$$\text{posisi item ke-j} = \sum_{i=1}^{10} y_{ij}, \quad y_{ij} \text{ mempunyai nilai dari } 0 \text{ sampai } 5$$

Perlu diingat bahwa untuk peserta, nilai j bergerak mulai dari 1 sampai 10. Sedangkan untuk item, nilai j bergerak dari 1 sampai 5.

Nilai rata-rata item ke-j = $\bar{y} = \frac{\sum_{i=1}^{10} y_{ij}}{10}$, y_{ij} *mempunyai nilai dari* 0 *ke* 5.

Nilai variansi item ke-j = $S^2_j = \frac{\sum_{i=1}^{10} (y_{ij} - \bar{y}_j)^2}{9}$

Korelasi dari item ke-k dan l = $r_{kl} = \frac{(\sum_{i=1}^{10} (y_{ik} - \bar{y}_k)(y_{il} - \bar{y}_l)) / 9}{S_k S_l}$

Untuk *variabel dikotomis* maka nilai dari item adalah 0 atau 1. Artinya item tersebut hanya mempunyai nilai benar atau salah. Kalau peserta ujian menjawab soal nomor k dengan benar maka ia akan memperoleh nilai 1. Sebaliknya kalau ia menjawab salah maka ia akan mendapat nilai 0. Bentuk data seperti ini akan mempunyai distribusi binomial. Bentuk mean dan variansinya akan dijabarkan dibawah ini.

Sebagai ilustrasi, misalkan ada N peserta ujian menjawab soal-soal sebanyak M. Anggap nilai N besar sekali, sehingga formula variansi nantinya cukup dibagi N saja (tidak perlu dibagi N-1). Dari sini akan diturunkan mean dan variansinya berdasarkan bentuk formula pada variabel nondikotomi.

$p_j = \bar{y}_j = \frac{\sum_{i=1}^{N} y_{ij}}{N}$, $y_{ij} = 0,1 \; dan \; j = 1,2,\ldots\ldots,M$

$$p_j = \frac{\text{banyaknya siswa yang menjawab benar pada item} - j}{N}$$

p_j adalah difinisi dari **derajat kesulitan** item-j. Dari sini didapat $q_j = 1 - p_j$.

$$S^2_j = \frac{\sum_{i=1}^{N} (y_{ij} - \bar{y}_j)^2}{N} = \frac{\sum y_{ij}^2 - 2\sum y_{ij}\bar{y}_j + \sum \bar{y}_j^2}{N}$$

Karena y_{ij} akan bernilai 0 atau 1 maka $y^2_{ij} = y_{ij}$, sehingga bentuk persamaan diatas akan menjadi:

$$S^2_j = \frac{\sum y_{ij} - 2\bar{y}_j \sum y_{ij} + \sum \bar{y}_j^2}{N} = \frac{\sum y_{ij}}{N} - 2\bar{y}_j \frac{\sum y_{ij}}{N} + \frac{\sum \bar{y}_j^2}{N}$$

$$S^2_j = \bar{y}_j - 2\bar{y}_j^2 + \bar{y}_j^2 = \bar{y}_j - \bar{y}_j^2 = \bar{y}_j(1-\bar{y}_j) = p_j q_j$$

Sehingga bentuk baru dari varinsi item ke-j : $\boxed{S^2_j = p_j q_j}$

Untuk korelasi antara item-k dan item-l *penurunan* formulanya adalah :

$$\rho_{kl} = \frac{\sum_{i=1}^{N}(X_{ik} - \bar{X}_k)(X_{il} - \bar{X}_l)/N}{\sigma_k \sigma_l}$$

$$\rho_{kl} = \frac{\sum (X_{ik}X_{il} - X_{ik}\bar{X}_l - \bar{X}_k X_{il} + \bar{X}_k \bar{X}_l)/N}{\sigma_k \sigma_l}$$

$$\rho_{kl} = \frac{\sum X_{ik}X_{il}/N - \bar{X}_k \bar{X}_l - \bar{X}_k \bar{X}_l + \bar{X}_k \bar{X}_l}{\sigma_k \sigma_l}$$

$$\rho_{kl} = \frac{\sum X_{ik}X_{il}/N - \bar{X}_k \bar{X}_l}{\sigma_k \sigma_l}$$

Karena $\bar{X}_k = p_k$ dan $\bar{X}_l = p_l$; $\sigma_k^2 = p_k q_k$ dan $\sigma_l^2 = p_l q_l$ maka persamaan diatas menjadi :

$$\boxed{\rho_{kl} = \frac{p_{kl} - p_k p_l}{\sqrt{p_k q_k p_l q_l}}}$$

dengan p_{kl} adalah proporsi bersama yang dimiliki oleh peserta ujian yang menjawab item-k dan item-l dengan benar.

Seseorang yang melakukan ujian diasumsikan mempunyai pengetahuan. Sewaktu menjawab soal ujian tersebut maka ada 3 kemungkinan yang dapat terjadi. Yang pertama ia tahu dan menjawab benar. Ke dua, ia tahu tetapi karena sesuatu hal ia menjawab salah. Sedangkan yang ke tiga adalah ia tidak tahu, tetapi kebetulan menjawab benar. Model matematika untuk ke tiga

kemungkinan di atas diturunkan oleh Spearman. Ia menuliskan argumen tentang "test score" secara psikologi dan matematik. Model skor Spearman adalah :

$$\mathbf{X = T + E}$$

dengan : X = skor yang diperoleh

T = skor yang sebenarnya (yang sesuai dengan pengetahuannya).

E = skor "kesalahan" dimana seseorang dapat memiliki ataupun kehilangan skor karena sesuatu hal.

Secara teori, kalau seorang siswa melakukan ujian berkali-kali tanpa batas, maka *nilai rata-rata* dari keseluruhan nilai ujian tersebut adalah nilai dia yang sebenarnya (T). Sedangkan nilai yang ia peroleh pada masing-masing ujian adalah nilai yang ia peroleh (X_i), dengan i =1, 2, ..N, dimana N membesar tak terbatas. Secara matematika "true score" dapat dijelaskan sebagai berikut.

Misal X adalah variabel random yang menunjukkan nilai seseorang pada satu mata pelajaran yang dilakukan berulang-ulang. Asumsi yang dimiliki adalah masing-masing ujian adalah ekivalen dan peserta ujian tidak dipengaruhi oleh masing-masing ujian tersebut. Misalkan X = { $x_1, x_2, \ldots$ } dengan x_i adalah nilai seseorang untuk ujian ke-i. Kemungkinan terjadinya x_i adalah p_i. Nilai "true score" didifinisikan sebagai nilai ekspektasi dari variabel random X. Variabel random X dapat juga berupa nilai-nilai peserta ujian untuk suatu mata pelajaran. Formulanya adalah :

$$T = \mu = \sum_{i=1}^{N} x_i p_i$$, i = 1, 2, ..., N ; dimana N tak terbatas.

Nilai diskrepansi (selisih) antara T dan X yaitu E, akan mempunyai nilai ekpektasi 0. *Buktinya* sebagai berikut :

$E = T - X$

Ekspektasi (E) = Ekspektasi (T - X)

Ekspektasi(E) = Ekspektasi (T) - Ekspektasi (X)

Ingat : Ekspektasi (T) = T ; dan Ekspektasi (X) = T; sehingga

Ekspektasi (E) = T - T = 0

IV.5 Sifat-sifat True Score dan Error Score

1. Rata-rata error score untuk populasi peserta ujian adalah 0.
2. Korelasi "true score" dan "error score" untuk populasi peserta ujian = 0.
3. Korelasi antar error scores = 0 ($\rho_{e_i e_j} = 0$, dengan $i \neq j$).

Dari konsep true score dan nilai error, tampak jelas bahwa dalam suatu ujian/penelitian maka nilai yang diperoleh adalah nilai observasi, bukannya nilai yang sebenarnya. Akan tetapi sorang guru atau peneliti akan lebih tertarik ke nilai sebenarnya. Untuk itu dicarilah *hubungan* dari nilai yang sebenarnya dengan nilai observasi. Caranya adalah melihat seberapa besarkah korelasi antara nilai observasi dan nilai sebenarnya dengan melakukan uji ber ulang ulang. Korelasi antara nilai sebenarnya (true scores) dengan nilai observasi (observed scores) dikenal dengan istilah *indeks reliabilitas* (reliability index).

Penurunan formulanya adalah :

$$X = T + E$$

$$\rho_{XT} = \frac{Cov(X,T)}{\sigma_X \sigma_T} = \frac{Cov((T+E),T)}{\sigma_X \sigma_T}$$

$$\rho_{XT} = \frac{\sum_{i=1}^{N}(T_i + E_i - \bar{T} + \bar{E})(T_i - \bar{T})}{N\sigma_X \sigma_T} =$$

$$\rho_{XT} = \frac{\sum(T_i^2 - E_i T_i - T_i \bar{T} - E_i \bar{T} - \bar{T} T_i - \bar{T}^2 - \bar{E} T_i - \bar{E} \bar{T})}{N\sigma_X \sigma_T}$$

$$\rho_{XT} = \frac{\sum_{i=1}^{N}(T_i^2 - \bar{T}^2) - \sum_{i=1}^{N}(E_i T_i - \bar{E} \bar{T})}{N\sigma_X \sigma_T}$$

karena $\sum_{i=1}^{N}(E_i T_i - \bar{E} \bar{T}) = 0$ *maka*

$$\rho_{XT} = \frac{\sum_{i=1}^{N}(T_i^2 - T^2)}{N\sigma_X \sigma_T} = \frac{\sigma_T^2}{\sigma_X \sigma_T} = \frac{\sigma_T}{\sigma_X}$$

Dengan ide yang sama, sekumpulan orang diuji dengan 2 macam uji yang paralel. Dengan uji paralel ini dapat dilihat seberapa besarkah reliabiliti dari suatu set alat uji .Dapat juga diartikan seberapa besarkah hubungan kemampuan/pengetahuan peserta-peserta pada ke dua set soal tersebut. Jangan lupa ke dua set tes tersebut adalah paralel.

Difinisi dua tes disebut paralel menurut Crocker & Algina (1986) adalah :

i) setiap peserta ujian mempunyai skor yang sama pada ke dua ujian tersebut.

ii) variansi error dari ke dua jenis tes tersebut sama.

Misal variabel random $X_1 = \{X_{11}, X_{12}, \ldots\ldots\}$ dan $X_2 = \{ X_{21}, X_{22}, \ldots\ldots \}$. Korelasi X_1dan X_2 adalah :

$$\rho_{X_1X_2} = \frac{\sum_{i=1}^{N}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{N\sigma_{x_1}\sigma_{x_2}}$$

$$\rho_{X_1X_2} = \frac{\sum_{i=1}^{N}(x_{1i}x_{2i} - \bar{x}_1\,\bar{x}_2)}{N\sigma_{x_1}\sigma_{x_2}}$$

Kalau seandainya tes tersebut betul-betul paralel maka nilai korelasi antara X_1 dan X_2 akan sama dengan 1.

IV.6 Indeks dan Koefisien Reliabilitas

Sudah jelas bahwa data yang dimiliki oleh peneliti adalah data observasi (observed scores) walaupun sebetulnya mereka lebih tertarik dengan data yang sebetulnya (true scores). Karena itu adalah relevan untuk bertanya "*seberapa dekatkah data observasi dengan data yang sebenarnya* ?". Salah satu cara untuk melihat hubungan mereka adalah dengan melihat korelasi dari kedua variabel ini. Koefisien korelasi yang menyatakan derajat hubungan antara nilai observasi dengan nilai yang sebetulnya dikenal dengan istilah ***indeks reliabilitas***.

Nilai observasi dapat dituliskan sebagai : X = T + E

Penghitung indeks reliabiliti akan *diturunkan* dibawah ini,

$$\rho_{XT} = \frac{Cov(X,T)}{\sqrt{\sigma_X^2\sigma_T^2}}$$

$$\rho_{XT} = \frac{\Sigma(X-\mu)(T-\mu)}{N\sigma_X\sigma_T}$$

$$\rho_{XT} = \frac{\Sigma(T+E-\mu)(T-\mu)}{N\sigma_X\sigma_T}$$

$$\rho_{XT} = \frac{\Sigma(T-\mu)^2 - E(T-\mu)}{N\sigma_X\sigma_T}$$

$$\rho_{XT} = \frac{\Sigma[(T-\mu)^2 - (E-0)(T-\mu)]}{N\sigma_X\sigma_T}$$

$$\rho_{XT} = \frac{\Sigma(T-\mu)^2}{N\sigma_X\sigma_T} + \frac{\Sigma(E-0)(T-\mu)}{N\sigma_X\sigma_T}$$

Karena E dan T saling bebas maka kovariansi antara E dan T adalah nol.

Sehingga :

$$\rho_{XT} = \frac{\Sigma(T-\mu)^2}{N\sigma_X\sigma_T} = \frac{\sigma_T^2}{\sigma_X\sigma_T}$$

$$\rho_{XT} = \frac{\sigma_T}{\sigma_X}$$

Dari sini tampak bahwa indeks reliabiliti dapat dinyatakan sebagai *ratio* dari *standard deviasi dari nilai yang sebenarnya* dengan *standard deviasi dari nilai observasi*. Tapi harus diingat bahwa ini adalah korelasi dari nilai yang sebenarnya dengan semua kemungkinan nilai-nilai observasi. Sehingga hal ini sulit untuk dipraktekkan karena nilai yang sebenarnya tidak dapat diambil secara langsung. Selain itu nilai observasi juga sulit didapat untuk semua kemungkinan yang ada.

Akan tetapi dengan ide tersebut dapat dilihat korelasi antara dua buah tes yang paralel. Dari sini akan muncul istilah *koefisien reliabiliti* yaitu korelasi antara skor-skor dari tes yang paralel.

IV.7 Standard Error of Measurement

Seperti yang sudah dijelaskan sebelumnya, taksiran nilai yang sebetulnya (true score) didapat dari sejumlah besar uji-uji yang paralel. Reliabiliti dalam hal ini adalah suatu cara untuk melihat proporsi variansi "true score" pada skor-skor yang didapat dari observasi. Tetapi dalam beberapa hal, para peneliti lebih tertarik pada "bagaimana pengukuran kesalahan mempengaruhi interpretasi skor perorangan". Untuk menjelaskan hal itu, pandang penjelasan berikut.

Misalkan X adalah variabel random yang berisikan nilai-nilai suatu pengukuran, dimana $X = \{x_1, x_2, \ldots\ldots\}$, dengan $x_i, i = 1,2,\ldots$ adalah nilai dari peserta ke-i. Masing-masing peserta ini kalau diuji berkali-kali akan membentuk suatu distribusi tersendiri, dimana nilai yang diperoleh diharapkan akan berkisar pada nilai yang sebetulnya (true score) dari masing-masing peserta.

Sekarang pandang $e_{ik} = x_{ik} - t_{ik}$, dimana $x_{ik} dan t_{ik}$ adalah nilai observasi dan nilai sebenarnya dari peserta ke-k. e_{ik} sendiri akan membentuk distribusi. Sebagai suatu distribusi, maka ia akan mempunyai standard deviasi. Kalau

standard deviasi ini dirata-ratakan maka hasilnya disebut dengan *standard error of measurement* dan disimbolkan dengan σ_E. *Penurunan* formulanya adalah :

X = T + E

Var (X) = Var (T+E)

Var (X) = Var (T) + Var (E) - Cov (T,E).

Karena T dan E saling bebas maka Cov (T,E) = 0.

Sehingga Var (E) = Var (X) - Var (T).

Bagi persamaan diatas dengan Var (X), sehingga didapat

Var (E)/Var (X) = Var (X)/Var (X) - Var (T)/Var (X)

Ingat bahwa Var (T)/Var (X) = $\rho_{TX} = \rho_{XX'}$ sehingga

Var (E)/Var (X) = 1 - $\rho_{XX'}$

Var (E) = (1- $\rho_{XX'}$) Var (X)

$$\sigma_E^2 = (1-\rho_{XX'})\sigma_X^2$$

atau $\sigma_E = \sigma_X\sqrt{1-\rho_{XX'}}$

Dari informasi σ_E ini didapatkan bahwa 68% dari *skor yang sebenarnya* akan berada pada interval $\bar{x}-\sigma_E < T < \bar{x}+\sigma_E$.

Seperti yang sudah dijelaskan sebelumnya, *nilai sebenarnya* dari peserta ujian didifinisikan sebagai *nilai rata-rata dari sejumlah besar tes yang paralel.* Akan tetapi sebenarnya ada difinisi lain dimana tes hanya dilakukan satu kali. Hanya disini tes terdiri dari banyak sekali soal-soal. Untuk jelasnya perhatikan ilustrasi berikut ini.

Pandang suatu himpunan A yang berisi soal-soal yang mempunyai nilai dikotomi. Dua atau lebih tes yang paralel dapat dibuat dari himpunan soal-soal ini dimana soal-soal dipilih secara random. Soal-soal pada tes tes tersebut tidak harus mempunyai mean ataupun variansi yang sama.

Nilai sebenarnya dari seseorang didifinisikan sebagai banyaknya soal-soal dari himpunan A yang dapat dijawab secara benar oleh orang tersebut. Banyak orang lebih senang mendifinisikan nilai sebenarnya sebagai proporsi soal-soal dari himpunan A yang dapat dijawab dengan benar oleh peserta tes. Dengan perkataan lain, nilai yang sebenarnya dari peserta-i adalah banyak soal yang dijawab (n) dikali dengan proporsi menjawab benar (P_i).

$$T_i = n\,P_i$$

Index of discrimination digunakan berdasarkan penentuan grup yang dianggap pandai dan grup yang tidak pandai. Untuk itu skor peserta ujian diurutkan terlebih dahulu. Setelah itu baru ditentukan batasan skor dari peserta yang dianggap pandai dan tidak pandai.

Menurut Kelley (1939), bilamana peserta ujian jumlahnya "besar", maka pembagian grup tidak perlu dibagi 2, tetapi cukup 27% bagian atas dari nilai tertinggi dan 27% bagian bawah dari nilai terendah Angka ini tidaklah baku, karena Beuchert dan Mendoza (1979) serta Englehart (1965) memakai 30% bagian atas untuk yang pandai serta 30% bagian bawah untuk yang tidak pandai.

Formula yang digunakan adalah : $D = p_u - p_l$
dimana p_u adalah proporsi dari dari grup atas yang menjawab soal dengan benar, dan p_l adalah proporsi dari grup bawah yang salah menjawab soal.

Ebel (1965) menawarkan kriteria peng"arti"an nilai D, sebagai berikut :

1. Jika $D \geq 40$ soal berfungsi dengan baik.
2. Jika $30 \leq D < 40$ soal tersebut harus direvisi sedikit.
3. Jika $20 \leq D < 30$ soal tersebut memerlukan revisi yang agak banyak
4. Jika $D < 20$ maka soal tersebut harus direvisi secara keseluruhan.

Point Biserial Correlation digunakan untuk set soal yang penilaiannya berdasarkan dikotomi (salah dapat 0, benar dapat 1). Formula yang dipakai adalah :

$$\rho_{pbis} = \frac{\mu_+ - \mu_X}{\sigma_X}\sqrt{\frac{p}{q}}$$

dimana: μ_+ *adalah nilai rata-rata responden yang menjawab benar pada soal tersebut.*
μ_+ *adalah nilai rata-rata secara keseluruhan.*
σ_x *adalah standard deviation*
p *adalah derajat kesulitan, dan* $q=1-p$

Menurut Crocker dan Algina (1986), bila jumlah soal tidak cukup "besar" maka nilai korelasi dapat membuat interpretasi yang menyesatkan, karena item skor berkontribusi pada skor total. Tetapi kalau jumlah soal cukup "besar" (25 soal, misalnya), maka problem tersebut tidak muncul. Untuk jumlah soal yang tidak besar, ada formula yang lebih cocok, yaitu :

$$\rho_{i(X-i)} = \frac{\rho_{X_i}\sigma_X - \sigma_i}{\sqrt{\sigma_i^2 + \sigma_X^2 - 2\rho_{X_i}\sigma_X\sigma_i}}$$

dimana: $\rho_{i(X-i)}$ *adalah korelasi antara skor item i dengan skor total dimana pengaruh item itu sudah dibuang.*

Biserial Correlation Coefficient digunakan bilamana distribusi skor peserta ujian berdistribusi normal. Statistik yang digunakan diturunkan mula-mula sekali oleh Pearson (1909) yang populer disebut "*biserial correlation coefficient*". Formulanya adalh :

$$\rho_{bis} = \frac{\mu_+ - \mu_X}{\sigma_X}(p / Y)$$

dimana : μ_+ *adalah nilai rata dari peserta yang menjawab benar*
μ_x *adalah nilai rata-rata semua peserta*
σ_x *adalah deviasi standar skor peserta ujian*
p adalah proporsi peserta ujian yang menjawab soal tersebut dengan benar
Y adalah nilai ordinat (skor z) untuk nilai p.

Dua koefisien yang terakhir adalah Phi Coefficient dan Tetrachoric Correlation Coefficient tidak dibahas pada penelitian ini.

V. Kesimpulan dan Saran

Item analisis adalah suatu cara atau alat yang digunakan untuk melihat "mutu" dari soal-soal yang diberikan kepada peserta ujian. Sudah tentu soal yang sama bila diberikan kepada beberapa kelompok siswa/mahasiswa yang berlainan dapat memberikan hasil atau "mutu" yang berbeda. Karena itu kelompok haruslah mempunyai kriteria yang sama agar hasil yang didapatpun akan sama.

Pada item analisis ada beberapa besaran statistik yang dipakai, dimana besaran-besaran tersebut mempunyai arti dan tujuan masing-masing. Besaran-besaran tersebut adalah p (derajat kesulitan), daya beda soal (point biserial), keandalan set soal (reliabilitas soal) - KR_{20}, nilai rata-rata siswa untuk satu set soal, dan deviasi standar skor siswa untuk satu set soal.

Derajat kesulitan menunjukkan seberapa "sulit" kah suatu soal yang dikerjakan oleh siswa. Derajat kesulitan bergerak antara 0 dan 1. Semakin besar nilai derajat kesulitan, semakin mudah soal tersebut.

Reliabilitas set soal adalah besaran yang menunjukkan seberapa konsistenkah suatu soal dijawab oleh siswa. Karena reliabilitas dihitung berdasarkan konsep korelasi, maka nilainya juga bergerak dari 0 ke 1. Semakin besar nilai reliabilitas tersebut semakin andal soal tersebut.

Daya beda soal digunakan untuk melihat kemampuan soal dalam membedakan si "pandai" dengan si "tidak pandai". Kalau memang soal tersebut baik maka ia akan mampu menyaring orang-orang yang pandai untuk menjawab benar dan orang-orang yang tidak pandai akan menjawab salah.

Nilai rata-rata dapat menunjukkan kemampuan rata-rata peserta ujian dalam menjawab set soal tersebut. Standar deviasinya menunjukkan seberapa besar sebaran kemampuan peserta ujian pada set soal itu.

Hampir semua formula yang digunakan untuk item analisis, formulanya diturunkan secara matematis. Beberapa formula memang tidak diturunkan, sebab adanya kemiripan cara pembuktian.

Untuk menjalankan item analisis sebenarnya ada beberapa kriteria yang harus dipenuhi. Kriteria tersebut antara lain adalah, soal-soal yang akan dianalisa haruslah diberikan kepada siswa/siswi yang sudah mempelajari materi yang akan diujikan. Kalau siswa/siswi ini tidak memenuhi kriteria tersebut maka dikhawatirkan kesimpulan dalam menganalisa soal akan "tidak benar". Kriteria yang harus dipenuhi tersebut memang tidak di jabarkan di penelitian ini. Selain itu, kalau misalnya kriteria yang diinginkan tidak dapat dipenuhi maka apa yang harus dilakukan agar item analisis tetap bisa dilaksanakan. Untuk menjawab pertanyaan dan persoalan tersebut maka diusulkan agar ada penelitian tersendiri untuk membahasnya.

VI. Daftar Pustaka

Beuchert, A.K and Mendoza, J.L (1979). A Monte Carlo comparison of ten item discrimination indices. Journal of Educational Measurement, 16, 109-118.

Ebel, R.L (1965). Measuring Educational Achievement. Englewood Cliffs, N.J: Prentice Hall.

Kelley, T.L (1939). Selection of upper and lower groups for the validation of test items. Journal of Educational Psychology, 30, 17-25.

Likert, R (1932). A technique for the measurement of attitudes. Archives of Psychology. 140., 44-53.

Linda Crocker dan James Algina (1986). Introduction to Classical & Modern Test Theory . Holt, Rinehart and Winston, Inc.

Lord, F.M and Novick, M.R (1968). Statistical theories of mental test scores. Reading Mass.: Addison-Wesley.

Pearson, K (1909). On a new method of determining a correlation between a measured character of A and a character of B, of which only the percentage of cases wherein B Exceeds (or fall short of) intensity is recorded for each grade of A. Biometrika, 7, 96-105.

Robert V Hogg & Allen T Craig (1978). Introduction to Mathematical Statistics. Macmillan Publishing Co., Inc: New York.

Spearman, C (1907). Demonstration of formulae for true measurement of correlation. American Journal of Psychology, 18, 161-169.

Stevens, SS (1946). On the Theory of Scales of Measurement. Science, 103, 265-275.

Torgerson, W.S (1958). Theory and Methods Of Scaling. John Willey: New York.

Weitzenhoffer, A.M (1951). Mathematical Structures and Psychological Measurement. Psychometrica, 16, 387-406.

Riwayat Hidup Peneliti

1. Nama : Dra. Nani Dianiyati

   Unit : Pusat Komputer/ Statistika FMIPA - UT

   Tempat / Tgl. Lahir : Ciamis / 25 September 1959

   Pendidikan : S1, Matematika, Universitas Pajajaran, 1984

   Pengalaman Penelitian : - Item Analisis, Statistika dan Penurunan Rumus Item Analisis, 1997.

2. Nama : Drs. Herman, MA

   Unit : Jurusan Statistika FMIPA - UT

   Tempat / Tgl. Lahir : Palembang / 25 Mei 1956

   Pendidikan : S1, Matematika, ITB, 1984
   S2, Educational Psychology, University of Victoria, 1993

   Pengalaman Penelitian : - A Study of Relationship between Achievement in Prerequisite Course in Applied Statistics and Economics Study Programs at Universitas Terbuka, 1993.

   - Uji Umur dan Penaksiran Keandalan Alat-alat yang Berdistribusi Eksponensial dengan Pendekatan Bayes, 1997.

   - Item Analisis, Statistika dan Penurunan Rumus Item Analisis, 1997.

Ucapan Terima Kasih

Sehubungan dengan selesainya penulisan laporan ini , penulis mengucapkan banyak terima kasih kepada pihak-pihak yang sudah membantu kelancaran penyelesaian penelitian ini.

Pertama-tama penulis mengucapkan terima kasih kepada Dekan FMIPA Universitas Terbuka, Dr. Djati Kerami yang sudah meluangkan waktunya dalam memberikan kritik dan saran untuk penelitian ini.

Selain itu juga penulis mengucapkan banyak terima kasih kepada Kepala Lembaga Penelitian, Dr. WBP Simanjuntak dan Kepala Pusat Studi Indonesia, Dr. Tian Belawati yang telah memberi kesempatan kepada penulis untuk mengikuti seleksi pembiayaan penelitian di Universitas Terbuka ini.

Tidak lupa juga penulis mengucapkan terima kasih kepada rekan-rekan yang sudah membantu penelitian ini, baik berupa kritik dan saran ataupun peminjaman buku-buku yang diperlukan.